Stout in Medical Science, Biology and Chemistry (SMSBC)
ISSN: 2979-2819
DOI: https://doi.org/10.61424/k0qrad59
Journal Homepage: https://stoutjournals.org/index.php/SMSBC

Risk-Calibrated Medical LLM Triage with Physician-Rubric Self-Verification and Patient-Facing Evidence Cards: An Empirical Health Bench Study

Oscar Liu
Computer Science, Brown University, Providence, RI, USA
Corresponding Author: Oscar Liu, **E-mail**: o_liu_research@yahoo.com

ARTICLE INFORMATION

Submitted: April 20, 2026
Accepted: July 27, 2026
Published: August 01, 2026
Volume: 2
Issue: 1

KEYWORDS

HealthBench; medical large language model; patient triage; calibration; selective prediction; retrieval-augmented generation; self-verification; patient safety; evidence card.

ABSTRACT

Medical large language models can answer patient questions fluently while remaining unreliable about urgency, uncertainty, and safe escalation. This study evaluated a risk-controlled post-generation architecture on HealthBench 2025, comprising 5,000 physician-annotated medical conversations, 57,237 weighted rubric criteria, 4,206 physician ideal responses, and 11,112 stored answers across 2,778 four-candidate cases. The architecture combined a word-level triage classifier, fold-local rubric retrieval, a word–character–retrieval ensemble with temperature scaling, deterministic answer reranking, triage-policy self-verification, selective escalation, and patient-facing evidence cards. Five-fold out-of-fold evaluation on 453 physician-labeled emergency-referral cases produced a macro-F1 of 0.645 for the self-verified ensemble versus 0.620 for direct TF-IDF; expected calibration error fell from 0.126 to 0.058, and under-triage fell from 18.5% to 16.1%. At a 0.70 confidence threshold, coverage was 32.0%, selective risk was 22.1%, and urgent-route recall reached 98.1%. On the 2,778 candidate-selection cases, rubric-RAG raised weighted rubric-alignment coverage from 0.801 to 0.845. Self-verification preserved 0.845 alignment while reducing the response-level triage-policy error proxy from 53.8% to 42.3%. The results show that retrieval improves content coverage, whereas a distinct safety check is needed to convert content quality into safer disposition behavior. Confidence-aware abstention and concise evidence cards provide an auditable interface for patient-facing use, while the remaining uncertainty supports clinician oversight rather than autonomous deployment.

1. Introduction

Large language models (LLMs) have moved rapidly from medical examination questions to patient communication, clinical documentation, and diagnostic support. Medical-domain studies show substantial gains in encoded knowledge and question answering (Singhal et al., 2023; Singhal et al., 2025), and patient-facing evaluations have reported strong ratings for response quality and empathy (Ayers et al., 2023). These advances do not by themselves establish safe triage. A response can be factually plausible, courteous, and complete while placing urgent referral language too late, omitting a red-flag condition, or expressing confidence that is poorly related to correctness. Realistic clinical decision studies continue to expose such gaps (Hager et al., 2024).

Triage is a particularly demanding test of medical AI because the cost of an error is asymmetric. Over-referral consumes scarce emergency capacity, but under-referral can delay time-sensitive care. Earlier symptom-checker audits therefore treated disposition advice as a separate outcome rather than a by-product of diagnostic accuracy (Semigran et al., 2015). The same principle applies to generative systems: answer quality, risk discrimination, confidence calibration, and the decision to defer should be measured independently. Metacognitive failure remains a documented weakness of medical LLMs (Griot et al., 2025), which makes an explicit verification layer more defensible than relying on the model's own verbal confidence.

HealthBench supplies a suitable empirical basis for this question. The benchmark was designed around realistic, frequently multi-turn health conversations and physician-authored, case-specific weighted rubrics (Arora et al., 2025). Its public release provides

5,000 conversations spanning communication, context seeking, emergency referrals, health-data tasks, complex responses, hedging, and global health (OpenAI, 2025a; OpenAI, 2025b). Unlike benchmarks that collapse quality into a single generic score, these rubrics make it possible to study which expected behaviors are covered, which harmful behaviors appear, and how performance changes across themes and evaluation axes.

This study asks whether a deterministic post-generation control system can improve both content selection and triage safety without generating new clinical prose. Four stored historical answers were available for 2,778 conversations. We compared a fixed direct-selection baseline with rubric-aware retrieval, then added an out-of-fold triage-policy check, calibrated confidence, and selective escalation. The selected answer was paired with a compact patient-facing evidence card that states the triage route, describes the verification process, and lists three matched physician-authored response checks from independent retrieval cases. This design separates the clinical answer from the control information needed to interpret and route it.

2. Literature Review

2.1 Medical foundation models and realistic evaluation

Medical foundation models are increasingly assessed across factuality, harm, bias, usefulness, and human preference rather than examination accuracy alone. Med-PaLM established a broad medical question-answering baseline (Singhal et al., 2023), while later work moved toward expert-level answers through instruction tuning, retrieval, and physician-based assessment (Singhal et al., 2025). Patient communication studies also distinguish quality from tone (Ayers et al., 2023). A literature-derived framework consequently organizes healthcare LLM review around quality, reasoning, expression, safety, and trust rather than one aggregate judgment (Tam et al., 2024).

Safety evaluation also needs failure-oriented measurement. Clinical decision-making experiments demonstrate that strong benchmark performance can degrade in realistic workflows (Hager et al., 2024). Medical summarization research separately measures hallucination severity and clinical safety (Asgari et al., 2025). HealthBench responds to these concerns with physician-written, instance-specific rubrics and deliberately diverse conversations (Arora et al., 2025). Its weighted criteria support granular analysis, but the official framework relies on a model grader; a transparent content-coverage proxy is therefore useful for examining deterministic reranking while preserving a clear distinction from the official score.

2.2 Calibration, selective prediction, and triage risk

A clinically useful confidence value must correspond to observed performance. Post-hoc temperature scaling remains a strong baseline for correcting overconfident neural probabilities (Guo et al., 2017), and language-model question answering has shown that calibration must be evaluated explicitly rather than inferred from likelihood alone (Jiang et al., 2021). Semantic uncertainty provides an additional signal for hallucination detection (Farquhar et al., 2024). In the present setting, calibration is operational: confidence determines whether a triage prediction is accepted or routed for review.

Selective classification formalizes this decision through the risk–coverage trade-off: a system answers fewer cases as its confidence threshold rises, and risk is computed among the retained cases (Geifman & El-Yaniv, 2017). Adaptive conformal classification offers finite-sample marginal coverage (Romano et al., 2020). The present experiments use fixed confidence thresholds rather than claim conformal coverage. This simpler design directly tests the clinically relevant routing question: when a low-confidence case is treated as an escalation, how much urgent-case recall is preserved?

2.3 Retrieval and self-verification

Retrieval-augmented generation grounds a model in external material at inference time (Lewis et al., 2020). Medicine-specific benchmarking shows that retrieval configuration and evaluation choices materially affect performance (Xiong et al., 2024). HealthBench permits a related but narrower strategy: retrieve physician-authored rubric patterns from independent Group 1 cases, then use those patterns to rank four already-written answers. Because no new clinical text is generated, changes can be attributed to selection rather than decoding.

Retrieval alone cannot guarantee safe behavior. Self-RAG combines retrieval with explicit reflection signals (Asai et al., 2024). Chain-of-verification separates an initial response from targeted checks (Dhuliawala et al., 2023). These approaches motivate the architecture tested here, but the safety signal is external to the selected answer: an out-of-fold triage model checks whether each candidate contains the referral behavior required by the predicted urgency class. This avoids treating fluent self-critique as proof of clinical reliability.

2.4 Medical VLMs, explanations, and patient-facing cards

Medicine is intrinsically multimodal. Med-PaLM M demonstrated a single model across clinical language, imaging, genomics, visual question answering, and radiology report generation (Tu et al., 2024). Work on Flamingo-CXR then showed that medical

VLM evaluation benefits from radiologist error correction, multiple clinical settings, and an assistive human–AI workflow (Tanno et al., 2025). These findings matter even for a text-only triage study. A future image explanation card must connect visual findings, textual recommendations, uncertainty, and referral action without implying that a plausible narrative validates the image interpretation.

Explanation design can influence perceived usefulness and trust in patient-facing healthcare systems (Z. Zhang et al., 2021), yet caution is essential. Healthcare explainability may create a false sense of validation (Ghassemi et al., 2021), and plausible explanations can obscure model failure (Babic et al., 2021). The evidence card in this study therefore describes process evidence—triage route, candidate comparison, and matched physician-authored response checks—rather than asserting that retrieved text is a clinical citation. This distinction aligns with governance principles that emphasize validity, reliability, transparency, and accountable risk management (Tabassi, 2023).

Recent medical-interface studies make this distinction concrete. Lightweight MedMNIST transfer pairs predictive performance with compact cross-modal representations (Mi et al., 2025). Subject-independent BCI research likewise treats calibration burden and cross-person generalization as deployment variables rather than secondary statistics (Wu et al., 2025). Earlier image-classification and hematology systems demonstrate the breadth of medical AI tasks to which this control problem applies, from cancer images to multiple-myeloma residual-disease assessment (J. Chen et al., 2024; Z. Zhong et al., 2020).

A parallel design literature has moved uncertainty and escalation cues into explicit card components. Radiology and breast-ultrasound cards visually separate the prediction, uncertainty, image cue, and warning state (Ye et al., 2025; Z. S. Zhong, Wu, & Mi, 2025). Patient-facing safety-card frameworks similarly emphasize rubric-based risk language and interface-level benchmarking (C. Li et al., 2025; B. Zhou et al., 2025). These precedents support the present card's narrow role: it displays routing and verification information, while the answer itself remains the object of content and safety evaluation.

2.5 Cross-domain calibration, imbalance, and selective decision control

The statistical problem is not unique to medicine. In credit-risk studies, probability calibration and uncertainty-driven rejection are evaluated separately from rank discrimination, while distribution-free uncertainty and model-risk explanations are treated as governance requirements (Y. Chen et al., 2023; Jin et al., 2024b). Extreme-imbalance experiments in fraud and bankruptcy prediction further show that nominal accuracy can conceal the class whose errors carry the highest cost (Y. Chen et al., 2024; Jin et al., 2024a). Fairness-aware counterfactual credit scoring extends the same principle to human review and actionable explanations (Xin, 2026b). These studies do not supply clinical evidence, but they justify reporting under-triage, urgent recall, calibration, and coverage as distinct outcomes.

Operational forecasting research reaches a similar conclusion under temporal and infrastructure shift. Calibrated foundation-model forecasts, cross-cloud transfer, few-shot cold starts, and workload-semantic risk curves all separate point performance from uncertainty under changed operating conditions (He et al., 2024; He et al., 2025; He et al., 2023; Zhao et al., 2025). Conformal demand envelopes and admission policies then convert uncertainty into bounded oversubscription or review decisions (S. Chen et al., 2024; Zhao et al., 2026). Probabilistic demand forecasting under weather and seasonal regime changes provides a further example of auditing interval coverage after shift (Xin, 2026e). Noisy-neighbor residual fusion and power-aware inventory planning illustrate why heterogeneous context must remain visible in the decision layer (He et al., 2026; Nie & Zheng, 2024).

The broader lesson is that uncertainty can be propagated through fusion rather than reported after a decision. Late-fusion calibration for 3D perception and macro-market uncertainty fusion provide cross-domain examples of combining signals with different failure modes (Xin, 2025b; H. Zhou & Zhang, 2025). Theoretical work on spectral rank aggregation and group synchronization additionally distinguishes estimator uncertainty from point recovery (Z. S. Zhong & Ling, 2024a; Z. S. Zhong & Ling, 2024b). Approximately shared-feature analysis and cross-dataset wearable transfer show why apparent source-domain success may not survive a target-domain change (J. Zhang, 2025; Z. S. Zhong, Pan, & Lei, 2025). Accordingly, the present findings are reported for HealthBench and not generalized to prospective clinical traffic.

2.6 Evidence retrieval, provenance, and rubric-grounded evaluation

Evidence-aware retrieval systems increasingly expose intermediate provenance rather than presenting a generated answer as an indivisible output. Cloud-native DevOps RAG and bilingual incident triage link retrieved operational evidence to answering and root-cause summaries (Liu et al., 2024; B. Zhang et al., 2024). Word-level evidence-chain methods make hallucination detection, source attribution, and deterministic provenance explicit (Jin, 2025a; C. Li et al., 2024). Although HealthBench rubrics are

evaluation criteria rather than clinical sources, these studies motivate retaining case identifiers, retrieved checks, and match values for every selection.

Coverage failures require a decision policy as well as a retriever. Budgeted multi-hop retrieval evaluates whether another hop is worth its cost (Su et al., 2025), while SQuAD 2.0 work explicitly calibrates abstention when retrieval does not support an answer (Z. S. Zhong, Chen, et al., 2025). Retrieval-quality prediction, accounting-aware due-diligence search, and numerical guardrails each test a different boundary between relevant text and a defensible conclusion (Y. Chen et al., 2025; Z. Li et al., 2026; R. Zhang et al., 2025). Scientific claim verification and multi-regulation legal RAG further demonstrate that evidence ranking must be aligned with the kind of claim being checked (J. Li & Zhou, 2026; Su, Chen, & Qian, 2026).

Retrieval also needs to work across language and presentation settings. Trust-calibrated multilingual RAG on humanitarian questions evaluates access and risk across languages rather than assuming that translation preserves evidence (Y. Chen & Xu, 2026). Offline reranking studies in code search, financial search, and review-grounded recommendation show that retrieval, ranking, and explanation faithfulness can be measured as separate stages (Chang et al., 2026; Y. Li, 2024; Meng et al., 2026). A rubric-grounded essay-scoring pipeline is especially relevant methodologically: it treats criterion coverage and auditable feedback as linked but nonidentical outputs (Xin, 2026a). This distinction informed the use of a transparent rubric-alignment proxy without relabeling it as the official HealthBench grader score.

2.7 Evidence cards as human-oversight interfaces

Card-based explanation is a cross-domain interface pattern, but its content must match the decision being supported. Scientific-search cards emphasize provenance and visual evidence hierarchy, accounting-review cards connect claims to disclosures, and recommendation cards expose the basis for ranking (Jin, 2025b; B. Zhang et al., 2025; K. Zhang et al., 2025). These designs are not interchangeable with medical advice. Their common value is structural: each separates a recommendation, its supporting evidence, uncertainty or limitations, and the action available to a human reviewer.

Human factors determine whether that structure helps. LLM-as-design-critic work compares generated feedback with human graphic-design judgment (Y. Li et al., 2025); capacity-dashboard briefs test how forecast risk becomes a concise visual decision object (Liu et al., 2025). Incident-visualization and privacy-risk cards likewise make source evidence, integrity concerns, and prompt-injection risk inspectable (Nie et al., 2026; Su, Rao, & Ma, 2026). The present card follows this restrained pattern by listing three matched checks and a route, not a persuasive narrative claiming clinical correctness.

Mental-health interfaces show why timing and context also matter. A therapist-facing copilot retrieves and ranks support from multi-turn counseling dialogues, whereas intake cards organize linked recommendations for review (Y. Zhang & H. Zhang, 2025a; Y. Zhang & H. Zhang, 2025b). Strategy-aware emotional-support response control separately evaluates whether the generated behavior matches an intended support strategy (Y. Zhang & Zhou, 2026). These precedents reinforce the decision to keep conversation context, selected content, and disposition behavior as separate analytic objects.

2.8 Longitudinal context, privacy, and reliability under attack

Longitudinal systems add a memory-governance problem to retrieval. Confidence-gated writing, time-aware retrieval, and explicit update or forgetting rules reduce the risk that stale user information silently becomes current evidence (Sun et al., 2023). Federated topic-preference learning adds a privacy-utility trade-off when personalization is distributed across users (Kuo et al., 2025). Privacy-safe aggregate modeling under identifier loss provides a cross-domain example of redesigning the measurement unit instead of reconstructing missing identities (Bai & Wu, 2026). HealthBench does not identify longitudinal patients, so this study does not construct persistent memory; it treats the visible turns of each conversation as a closed context.

Reliability must also survive deliberate manipulation. Continual red-teaming targets emerging jailbreaks, while multi-stage refusal and evidence-based self-verification aim to preserve useful responses under adversarial prompts (Zheng & Li, 2024; Zheng et al., 2023; Zheng et al., 2024). CloudTrail attack-chain modeling demonstrates the value of retaining event order and interpretable stage transitions (Bai et al., 2026). These security studies motivate a general design rule for medical triage: verification should inspect behavior and evidence, not merely accept a fluent self-description of safety.

Intrusion and anomaly detection provide useful stress tests for the same layered architecture. TinyLLM and language-guided network detectors combine compact models with operational features (Y. Li & Lu, 2025; Lu & Zhou, 2024); autoencoder-based IoT classification and predictive DDoS mitigation illustrate the class-imbalance and rapid-routing setting (Wang et al., 2024; Xin et al., 2024). Host-based sequence models then add suspicious-window localization and evidence-grounded forensic narratives (Xin, 2026c). The analogy is methodological rather than clinical: a high-risk event must be detected, localized, explained, and escalated with each stage evaluated independently.

Log and agent studies make the audit trail still more explicit. Multi-source root-cause attribution, retrieved normal prototypes, and self-supervised log models compare anomaly scores with evidence useful for diagnosis (Liu et al., 2023; Xin, 2025a; Xin, 2026f). Conformal alert control and retrieval-grounded HDFS narratives add abstention and deterministic evidence generation (Sun et al., 2026; Xin, 2026d). Cost-aware AIOps routing and generalist-agent trajectory prediction further separate task success, failure detection, and escalation cost (C. Li et al., 2026; Z. S. Zhong et al., 2026). Together, these precedents support the manuscript's layered evaluation without implying that infrastructure labels are substitutes for physician judgment.

3. Methodology

3.1 Dataset and analytic cohorts

The analysis used the May 7, 2025 HealthBench JSONL release. Integrity checks confirmed 60,258,154 bytes, SHA-256 e99dd3c6372c10d6fcc5e385c5fae69d0dd40392dae56836ef9493ae324ecd2f, and 5,000 valid records. Group 1 provides physician ideals without stored model answers; Groups 2 and 3 provide four historical answers per case. Answer positions lack model labels, so all comparisons are candidate-level.

Table 1. Dataset composition and experimental availability

Measure	Count	Analytic role
Conversations	5,000	All analyses
Single-message conversations	2,915	Dataset profile
Multi-turn conversations	2,085	Dataset profile and subgroups
Physician rubric criteria	57,237	Content and safety measurement
Positive / negative criteria	39,662 / 17,575	Rubric polarity
Physician ideal completions	4,206	Proxy calibration and similarity
Four-answer conversations	2,778	Response-selection cohort
Stored historical answers	11,112	Four candidates per response case
Physician-labeled triage cases	453	Five-fold out-of-fold triage

Note. Counts were computed from the complete 5,000-record JSONL file. The response cohort consists of Groups 2 and 3; Group 1 supplies independent retrieval material.

Table 2. HealthBench themes, annotations, and stored-answer coverage

Theme	N	%	Ideals	4 answers	Mean turns	Median rubrics
Global Health	1,097	21.94 %	914	601	2.15	12
Hedging	1,071	21.42 %	939	612	2.83	12
Communication	919	18.38 %	687	460	2.88	10
Context Seeking	594	11.88 %	552	364	3.12	13
Emergency Referrals	482	9.64%	329	224	2.08	11
Health Data Tasks	477	9.54%	449	296	1.85	10
Complex Responses	360	7.20%	336	221	3.57	10

Note. Theme percentages sum to 100%. Four-answer counts identify cases eligible for deterministic candidate selection.

The file contains 57,237 rubric instances: 39,662 positive and 17,575 negative. Among 482 emergency-referral conversations, 453 had physician categories—134 non-emergent, 180 conditionally emergent, and 139 emergent—and formed the triage cohort. The four-answer cohort contained 208 labeled emergency cases for response-level safety analysis.

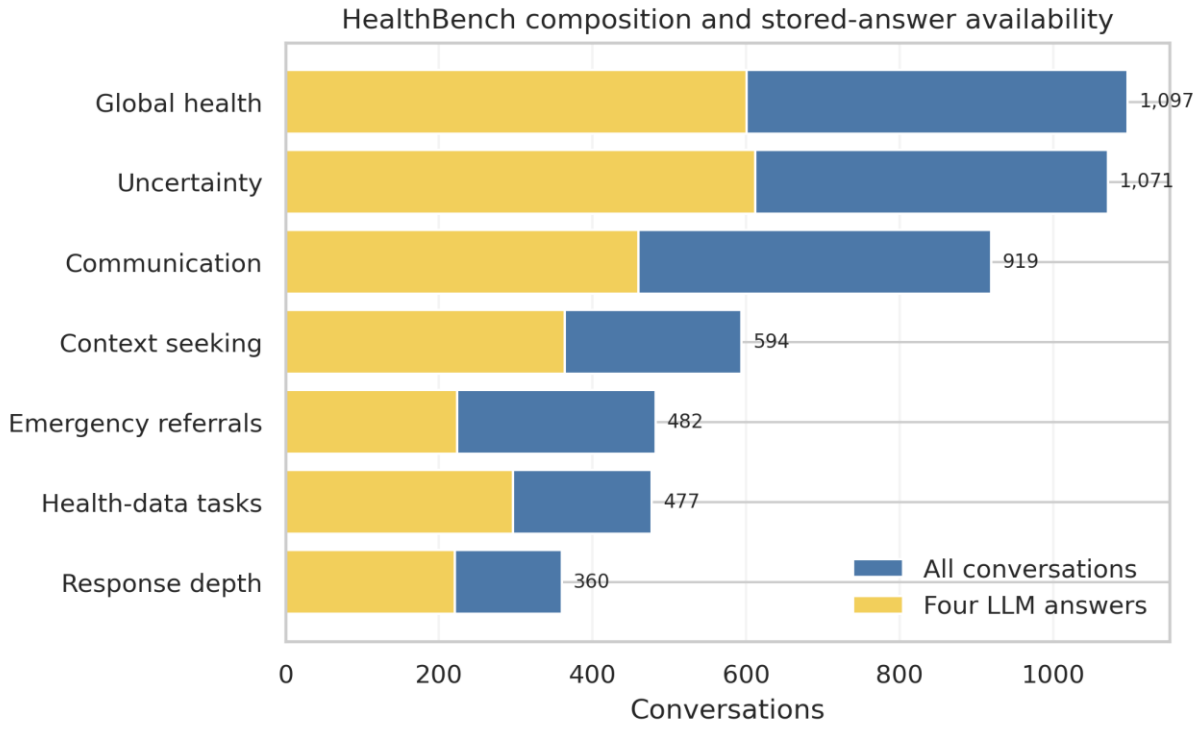

Figure 1. Distribution of conversations and response-selection availability across the seven HealthBench themes. The display labels "Uncertainty" and "Response depth" correspond to Table 2's "Hedging" and "Complex Responses," respectively.

3.2 Experimental architecture

Figure 2 shows the control path. Five-fold out-of-fold triage preceded retrieval of eight Group 1 cases. Their positive rubrics ranked four stored answers by coverage, similarity, relevance, consensus, and length. Self-verification added triage-policy compatibility; low confidence triggered abstention, and each selected answer received a separate evidence card.

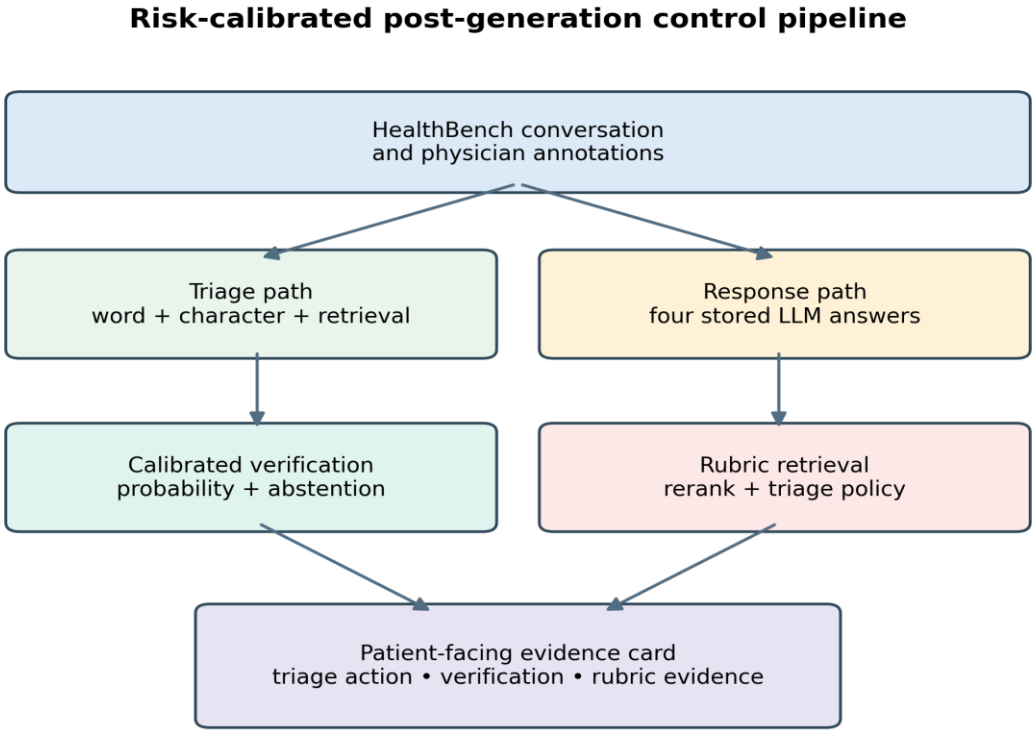

Figure 2. Risk-calibrated post-generation control pipeline used for triage, answer selection, abstention, and evidence-card rendering.

Table 3. Experimental systems and leakage controls

Component	Implementation	Control
Direct triage	Word 1-2-gram TF-IDF logistic regression	5-fold OOF
Rubric-RAG triage	Cosine-weighted 15-neighbor retrieval	Fold-local index
Self-verified triage	Word + character + retrieval mean; temperature scaled	Fold-local calibration
Selective escalation	Abstain below fixed confidence thresholds 0.50-0.90	No test tuning
Direct stored answer	SHA256(prompt_id) modulo four historical LLM answers	2,778 conversations
Rubric-RAG reranker	Eight-neighbor Group 1 rubric retrieval and candidate scoring	Target rubric excluded
Self-verified reranker	Retrieval score + OOF triage-policy check + length regularizer	No new clinical text
Evidence card	Selected answer plus top-three matched retrieved rubric items	Behavioral evidence

Note. OOF denotes out-of-fold. Target-case rubrics were used only for evaluation and never for retrieval-based candidate selection.

3.3 Triage models, calibration, and selective escalation

The direct triage model was a class-weighted multinomial logistic regression with word one- and two-gram TF-IDF (20,000 features, C = 2.0). The retrieval baseline used cosine-weighted 15-nearest neighbors with 0.03 similarity smoothing. The ensemble averaged the word model, a character three- to five-gram logistic regression (30,000 features, C = 1.5), and retrieval probabilities.

Five stratified folds used seed 20,250,729. Within each training fold, 20% formed a calibration subset. Negative-log-likelihood minimization selected a temperature for the untouched fold. Mean fold temperature was 0.471 (SD = 0.033); every downstream triage-policy check was therefore out of fold.

Selective escalation used maximum calibrated probability at fixed thresholds 0.50–0.90. Below-threshold cases were routed for review; accepted conditional or emergent predictions were routed urgently. Coverage measured acceptance, risk measured accepted-case error, and urgent-route recall credited urgent predictions and abstentions. Test folds did not tune thresholds.

3.4 Rubric retrieval, candidate selection, and evidence cards

The response experiment included all 2,778 conversations with four stored answers: 1,485 Group 2 and 1,293 Group 3 cases. Direct selection used the full prompt-identifier SHA-256 integer modulo four. Rubric-RAG fitted 30,000-feature word TF-IDF retrieval on 1,428 Group 1 prompts; eight neighbors supplied positive criteria. Target rubrics and physician ideals never entered selection.

Candidates and criteria used L2-normalized, fit-free 2^14-dimensional word and 2^14-dimensional character hashing features. A Group 1 calibration procedure compared every positive rubric with its own physician ideal and a same-theme mismatched ideal. Among 12,010 matched and 12,010 mismatched pairs, the selected similarity threshold was 0.1116, producing F1 = 0.678 and balanced accuracy = 0.641. Rubric coverage was positive rubric weight meeting that threshold.

Rubric-RAG combined coverage (0.45), criterion similarity (0.15), prompt relevance (0.15), candidate consensus (0.15), and excess length (−0.05). Self-verification added 0.20 for out-of-fold triage-policy compatibility and 0.025 for median-length fit. Emergent cases required urgent language in the first three sentences; conditional cases required urgent and conditional language; non-emergent cases penalized unconditional early emergency advice.

Evidence cards ranked retrieved criteria by 0.8 answer similarity plus 0.2 normalized rubric weight, deduplicated them, and retained three. Each card stated the triage route, four-candidate/eight-case verification, matched checks, and a safety note. All 2,778 JSONL records retained source identifiers and match values. Content metrics used the unchanged clinical answer; card length was separate.

3.5 Outcomes and statistical analysis

Table 4. Primary outcomes, operational definitions, and denominators

Outcome	Operational definition	N
Triage macro-F1	Unweighted mean F1 across non-emergent, conditional, and emergent classes	453
Urgent recall	Recall after combining conditional and emergent cases as urgent	319
Under-triage	Fraction of 453 cases with predicted ordinal severity below the physician label	453
ECE	Ten equal-frequency bins of confidence versus observed correctness	453 / 1,293
Multiclass Brier	Mean unhalved sum of squared error across three classes; range 0–2	453
Rubric alignment	Positive-rubric-weighted hashed similarity coverage at the calibrated threshold	2,778
Physician-ideal similarity	Cosine similarity between hashed answer and physician ideal representations	2,778
Triage-policy error	Rule-based failure to express the physician-labeled referral behavior	208
Evidence top-3 coverage	Fraction of the three highest-weight target positive rubrics covered by the answer	2,778

Note. Response content measures are deterministic lexical-semantic proxies. They are not the official HealthBench model-grader score and do not substitute for clinician judgment.

Table 4 defines all outcomes. Physician-ideal similarity is hashed word–character cosine rather than a clinical semantic grade. Group 2 trained response confidence to predict whether selected-answer similarity reached the within-case candidate median; Group 3 tested it. This measures relative selection, not absolute medical correctness.

Paired response and triage differences used 1,000 conversation-level bootstrap resamples; discordant triage correctness also used an exact McNemar comparison. Fixed selections supported theme, axis, group, turn, and ablation analyses. Timed blocks covered triage cross-validation and response reranking only.

4. Results and Discussion

4.1 Triage discrimination and calibration

Table 5. Five-fold out-of-fold triage performance on 453 physician-labeled cases

Method	Acc.	Bal. acc.	Macro-F1	Urgent rec.	Under-triage	ECE	Brier	NLL
Direct TF-IDF	0.614	0.620	0.620	0.859	18.5%	0.126	0.525	0.887
Rubric-RAG kNN	0.583	0.573	0.578	0.875	18.5%	0.082	0.530	0.951
Self-verified ensemble	0.642	0.641	0.645	0.881	16.1%	0.058	0.480	0.802

Note. Lower values are better for under-triage, ECE, Brier, and NLL. Brier is the unhalved three-class score.

The self-verified ensemble led on accuracy 0.642, balanced accuracy 0.641, and macro-F1 0.645. Relative to direct TF-IDF, macro-F1 increased by 0.025 (95% CI −0.003 to 0.054). The ensemble corrected 31 direct errors and introduced 18 (exact $p = 0.085$); the point estimate favors the ensemble, but 453 cases do not resolve a small advantage conclusively.

Safety-oriented metrics moved in the same direction. Urgent recall increased from 85.9% to 88.1%, and under-triage decreased from 18.5% to 16.1%. kNN raised urgent recall but lowered macro-F1 and non-emergent recall, favoring safer but less specific escalation. The ensemble balanced that tendency.

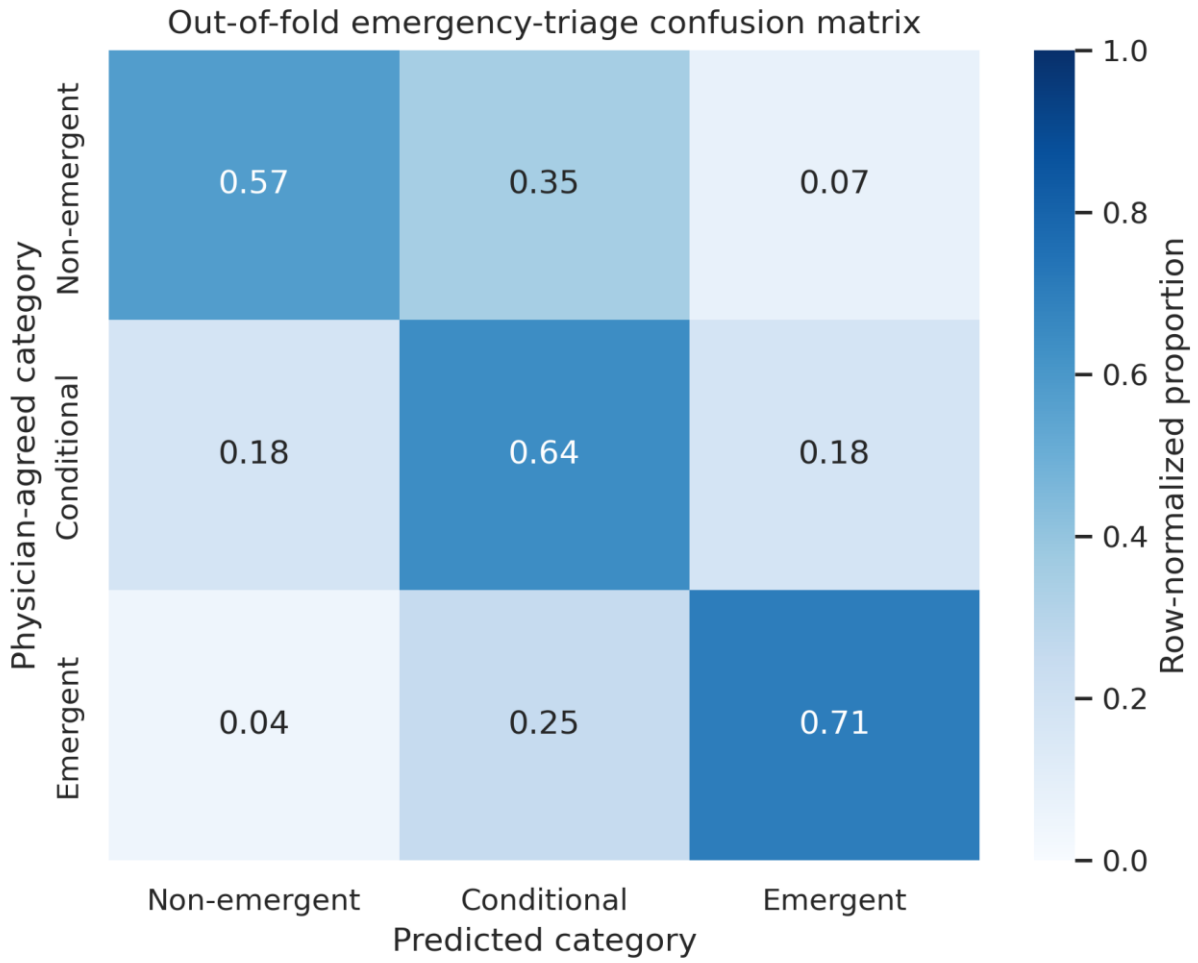

Figure 3. Out-of-fold row-normalized confusion matrix for the self-verified ensemble.

Conditional urgency was hardest because its language overlaps reassurance and immediate escalation. The ensemble recovered 70.5% of emergent cases and 64.4% of conditional cases, compared with 67.6% and 55.0% for direct TF-IDF. Its non-emergent recall was below the direct model's 63.4%, consistent with shifting marginal cases upward.

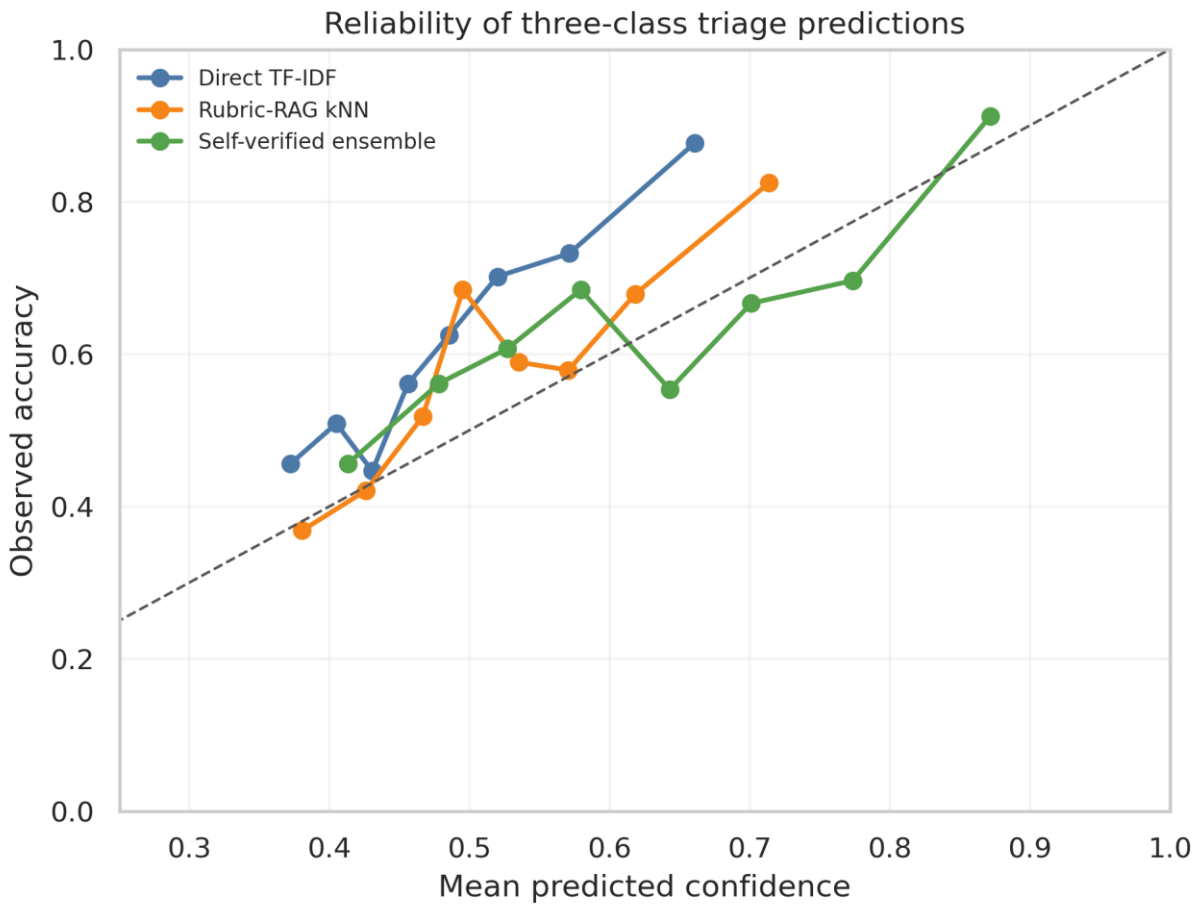

Figure 4. Reliability curves for the three triage systems using ten equal-frequency confidence bins.

Calibration improved more clearly than discrimination. ECE fell from 0.126 for direct TF-IDF and 0.082 for kNN to 0.058 for the scaled ensemble; Brier score and negative log likelihood were also lowest. Figure 4 shows smaller confidence–accuracy gaps after temperature scaling.

Table 6. Selective escalation for the calibrated self-verified triage ensemble

Threshold	Coverage	Selective risk	Accepted	Abstained	Urgent-route recall
0.50	77.0%	31.2%	349	104	92.8%
0.65	42.6%	26.4%	193	260	96.2%
0.70	32.0%	22.1%	145	308	98.1%
0.80	15.5%	12.9%	70	383	99.4%
0.90	2.6%	0.0%	12	441	100.0%

Note. Abstained cases were counted as routed for review. Selective risk is the error rate among accepted cases.

Selective escalation exposed the operating trade-off. At 0.50, 77.0% of cases were accepted with 31.2% selective risk and 92.8% urgent-route recall. At 0.70, coverage fell to 32.0%, risk to 22.1%, and urgent-route recall rose to 98.1%. At 0.80, only 15.5% were accepted, with 12.9% risk and 99.4% urgent-route recall. At 0.90, 12 cases were accepted without an observed error and every urgent case was routed; this small endpoint does not establish zero future risk.

4.2 Response selection and safety behavior

Table 7. Response-selection outcomes on 2,778 four-candidate conversations

Method	Rubric align.	Ideal sim.	Policy error	Emergent rec.	Conditional rec.	Top-3 cov.	Words
Direct stored answer	0.801	0.642	53.8%	32.3%	23.1%	0.824	298.8
Rubric-RAG reranker	0.845	0.676	54.8%	33.9%	19.8%	0.861	304.9
Self-verified reranker	0.845	0.675	42.3%	56.5%	33.0%	0.861	304.0
Verified + evidence card	0.845	0.675	42.3%	56.5%	33.0%	0.861	406.3

Note. Content metrics use the selected clinical answer. For the evidence-card method, Words includes the displayed card; the answer itself is identical to the self-verified selection.

Rubric-RAG produced the highest overall rubric-alignment point estimate (0.845) and physician-ideal similarity (0.676), compared with 0.801 and 0.642 for direct selection. Mean length rose by fewer than ten words before the card. Retrieval increased coverage of the three highest-weight target rubrics from 0.824 to 0.861.

Retrieval alone did not improve referral safety. Its triage-policy error proxy was 54.8%, slightly above the direct baseline's 53.8%. Adding the out-of-fold policy check reduced the error to 42.3% while preserving a rubric-alignment score of 0.845. Emergent early-referral recall rose to 56.5% versus 32.3% directly, and conditional recall exceeded both baselines. Retrieval selected better content; the triage constraint selected safer disposition behavior.

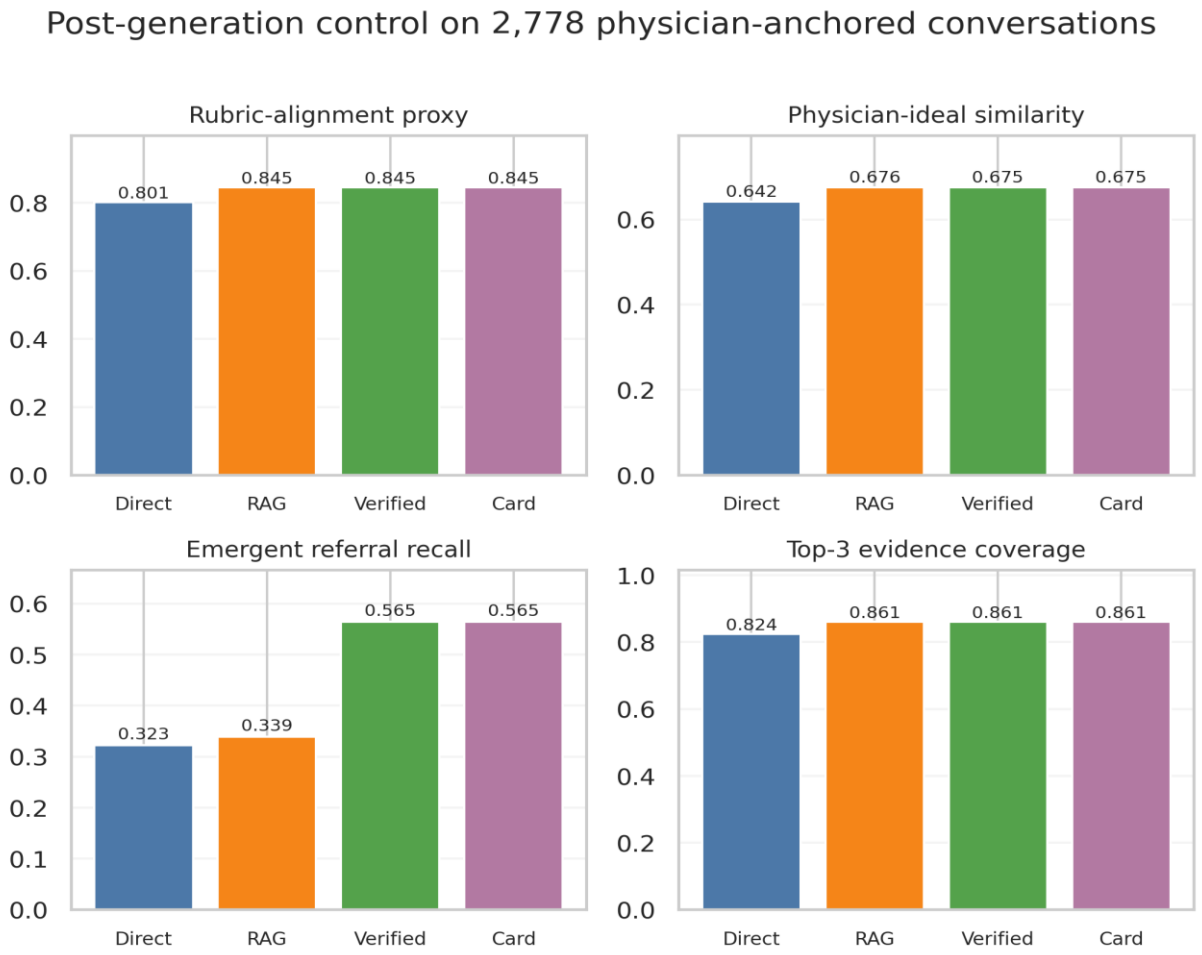

Figure 5. Rubric alignment, physician-ideal similarity, emergent-referral recall, and top-three evidence coverage across response-selection methods.

Table 8. Paired response differences with 1,000-bootstrap 95% confidence intervals

Comparison	Metric	Mean difference	95% CI	Paired N
RAG − Direct	Rubric alignment	0.044	[0.039, 0.049]	2,778
RAG − Direct	Ideal similarity	0.033	[0.026, 0.040]	2,778
Verified − Direct	Rubric alignment	0.043	[0.038, 0.049]	2,778
Verified − Direct	Ideal similarity	0.033	[0.025, 0.041]	2,778
Verified − RAG	Rubric alignment	-0.001	[-0.002, 0.000]	2,778
Verified − RAG	Ideal similarity	-0.000	[-0.002, 0.002]	2,778
RAG − Direct	Policy error	0.010	[-0.038, 0.058]	208
Verified − Direct	Policy error	-0.115	[-0.168, -0.062]	208
Verified − RAG	Policy error	-0.125	[-0.168, -0.082]	208

Note. Positive differences favor the left method for alignment and similarity; negative differences favor it for policy error.

Bootstrap inference separated the robust effects from near ties. RAG exceeded direct selection in rubric alignment by 0.044 (95% CI 0.039 to 0.049) and ideal similarity by 0.033 (0.026 to 0.040). Self-verification and RAG did not differ clearly on content, but self-verification reduced policy error by 0.125 (95% CI 0.082 to 0.168 in absolute magnitude) without a material content loss.

Table 9. Group 2-to-Group 3 calibration of relative candidate success

Method	Group 2	Group 3	Accuracy	Brier	ECE	NLL
Direct stored answer	1,485	1,293	0.553	0.261	0.126	0.719
Rubric-RAG reranker	1,485	1,293	0.614	0.239	0.065	0.673
Self-verified reranker	1,485	1,293	0.609	0.242	0.074	0.679
Verified + evidence card	1,485	1,293	0.609	0.242	0.074	0.679

Note. The binary target indicates whether selected-answer ideal similarity reached the within-case median among four candidates.

RAG had the highest Group 3 accuracy and lowest Brier, ECE, and NLL. Self-verification remained better calibrated than direct selection but trailed RAG because its policy check targets referral behavior, whereas the confidence target uses ideal-text similarity. Quality and disposition safety need separate probabilities.

4.3 Theme, axis, and subgroup performance

Table 10. Self-verified response performance by HealthBench theme

Theme	N	Rubric alignment	Ideal similarity	Top-3 coverage
Communication	460	0.900	0.686	0.904
Hedging	612	0.810	0.672	0.850
Emergency Referrals	224	0.909	0.658	0.924
Global Health	601	0.752	0.681	0.759
Health Data Tasks	296	0.902	0.655	0.917
Complex Responses	221	0.859	0.692	0.885

Theme	N	Rubric alignment	Ideal similarity	Top-3 coverage
Context Seeking	364	0.891	0.676	0.892

Note. N is the number of four-answer conversations in each theme.

Theme analysis identified global health as the hardest content domain: self-verified rubric alignment was 0.752 and top-three coverage 0.759. Hedging was next lowest at 0.810. Four themes exceeded 0.890. Variation in resources and care pathways may weaken eight-case lexical transfer, cautioning against geographically uniform safety claims.

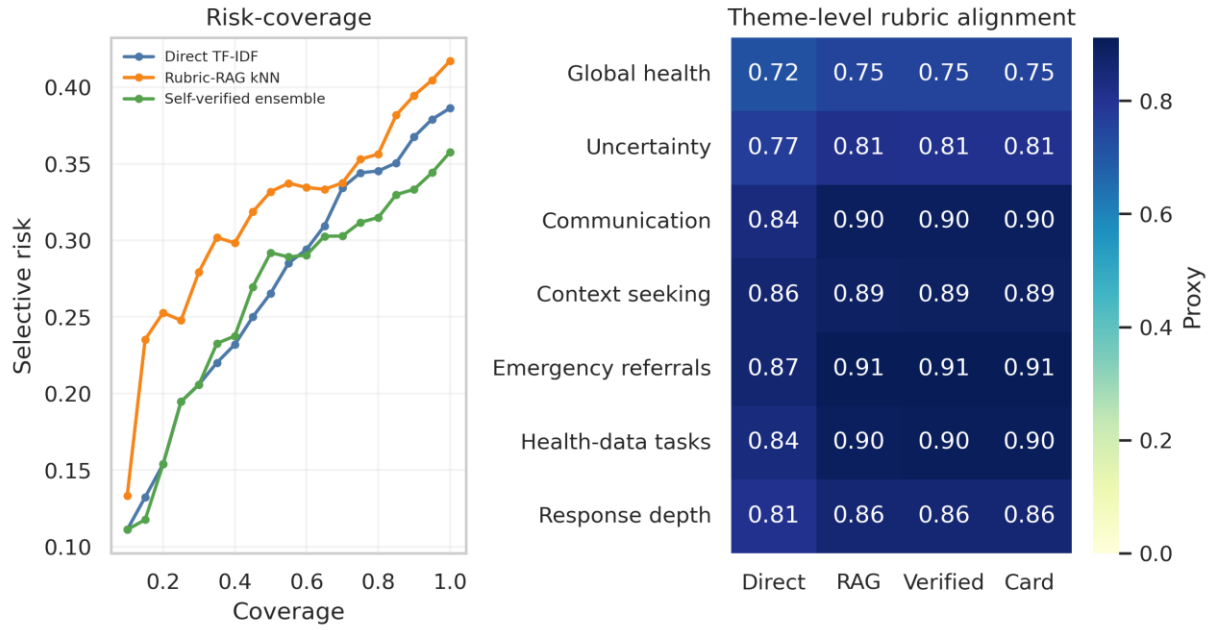

Figure 6. Selective triage risk–coverage relationship and theme-level rubric-alignment heat map.

Table 11. Positive-rubric coverage by evaluation axis

Axis	Valid N	Direct	Rubric-RAG	Self-verified
Accuracy	2,342	0.820	0.861	0.861
Completeness	2,432	0.821	0.857	0.855
Context Awareness	1,916	0.821	0.880	0.881
Communication Quality	1,248	0.676	0.742	0.738
Instruction Following	598	0.743	0.821	0.822

Note. Valid N varies because not every conversation contains a positive rubric for every axis.

Retrieval improved all five axes. The self-verified system covered 0.861 of accuracy weight, 0.855 of completeness weight, 0.881 of context-awareness weight, 0.738 of communication-quality weight, and 0.822 of instruction-following weight. Communication quality remained weakest despite strong theme-level performance, indicating persistent style, empathy, or clarity gaps.

Table 12. Self-verified response outcomes by completion group and conversation length

Group	Conversation	N	Rubric alignment	Ideal similarity	Policy error
Group 3	Single-message	769	0.874	0.659	34.9%
Group 2	Multi-turn	653	0.798	0.681	48.8%
Group 3	Multi-turn	524	0.831	0.632	25.6%

Group	Conversation	N	Rubric alignment	Ideal similarity	Policy error
Group 2	Single-message	832	0.863	0.713	55.6%

Note. Policy-error denominators include only physician-labeled emergency-referral cases within each subgroup.

Group 2 single-message cases had the highest physician-ideal similarity (0.713), whereas Group 3 multi-turn cases had the lowest (0.632). Rubric alignment was also lower for multi-turn conversations in both groups. Policy error used smaller emergency-only denominators and did not track content monotonically. Multi-turn state tracking and separate safety evaluation remain necessary.

4.4 Ablation, card production, and practical interpretation

Table 13. Response-selection ablation analysis

Configuration	Rubric align.	Ideal sim.	Policy error	Top-3 cov.	Words
Full self-verification	0.845	0.675	42.3%	0.861	304.0
No triage-policy check	0.845	0.676	54.8%	0.861	304.9
No rubric retrieval	0.801	0.642	53.8%	0.824	298.8
No length regularizer	0.845	0.676	42.3%	0.861	304.6
No candidate consensus	0.846	0.673	42.3%	0.862	299.0

Note. Each row used the same 2,778 conversations. Removing rubric retrieval reverts to the deterministic direct-selection rule.

The ablation results locate the main mechanisms. Removing the triage-policy check returned policy error to the RAG level while leaving content scores essentially unchanged. Removing rubric retrieval caused the largest content loss and worsened policy error. Length and consensus had smaller effects. The system therefore depends on retrieval for coverage and the policy check for safety.

Table 14. Timed stages and patient-facing evidence-card output

Stage	Examples	Total time (s)	Per-example (ms) / output
Response reranking	2,778	240.14	86.44
Triage cross-validation	453	19.55	43.15
Evidence-card rendering	2,778	Included in reranking	102.3 mean card words
Structured card output	2,778	2,778 JSONL records	3.0 checks per card

Note. Timing excludes data loading, threshold calibration, retrieval-index construction, file writing, and plotting.

Triage cross-validation required 19.55 seconds, and response reranking required 240.14 seconds. The pipeline generated exactly 2,778 evidence cards, each with three matched checks; their mean added length was 102.3 words. Cards changed display length, not the selected answer or its scores, so routing and verification remain separately inspectable.

The empirical findings support a layered safety architecture rather than a single composite score. Content retrieval, triage discrimination, probability calibration, answer-policy checking, abstention, and explanation each solved a different problem. None established clinical validity. Hashed text similarities can miss contradictions, dose errors, and unsupported claims; referral rules can miss paraphrases or reward formulaic warnings. The 208-case emergency response subset also yields wider uncertainty than content results.

The stored answers also bound the interpretation. The experiment selects among four historical candidates and does not compare current foundation models by name. It tests fixed-set post-generation control, not a newly trained model. All groups share one benchmark distribution. External validation should add prospective messages, clinician adjudication, diverse care pathways, and delayed-harm outcomes.

The VLM extension is conceptually direct but empirically separate. A radiology or dermatology system could place image regions, report statements, uncertainty, and triage route on an imaging explanation card. Visual grounding still requires image-level experiments and expert localization. These text results support the control logic without claiming image-interpretation performance.

5. Conclusions

A full empirical evaluation of HealthBench 2025 showed that deterministic post-generation controls can improve distinct dimensions of medical LLM behavior. Rubric retrieval raised weighted content coverage and similarity to physician ideals across 2,778 four-answer conversations. An out-of-fold triage-policy check preserved those content gains while substantially reducing unsafe referral behavior in the physician-labeled emergency subset. The temperature-scaled triage ensemble produced the best discrimination, calibration, and under-triage point estimates among the tested systems, and confidence-based abstention exposed a clear risk–coverage trade-off. Patient-facing evidence cards made the selected route and verification basis inspectable for every response.

The results also define the remaining work. The macro-F1 advantage over direct triage had an interval that included zero, communication-quality coverage remained weakest, and global-health conversations formed the hardest theme. Text-similarity and rule-based safety measures require clinician validation, and a card must not be mistaken for proof that the underlying advice is correct. Accordingly, the strongest use case is a clinician-supervised system that retrieves relevant quality criteria, checks disposition behavior, abstains when confidence is low, and presents a concise evidence card. This architecture is also a practical template for future medical VLM studies in which image findings and patient-facing explanations must be linked to calibrated referral decisions.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers

Artificial Intelligence (AI) Use Disclosure: "The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Arora, R. K., Wei, J., Soskin Hicks, R., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., & Singhal, K. (2025). HealthBench: Evaluating large language models towards improved human health [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.08775

[2] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=hSyW5go0v8

[3] Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Au Yeung, J., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine, 8*(1), Article 274. https://doi.org/10.1038/s41746-025-01670-7

[4] Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine, 183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

[5] Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science, 373*(6552), 284–286. https://doi.org/10.1126/science.abg1834

[6] Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. *International Journal of Electronics and Communications Systems, 6*(1), 83–95. https://doi.org/10.24042/ijecs.v6i1.30533

[7] Bai, J., Chen, S., Zheng, D., & Kuo, M.-J. (2026). Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing. *Information, Electrical and Electronics Engineering, 6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923

[8] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon Reviews. *Journal of Electrical Engineering and Computer Science, 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

[9] Chen, J., Xiong, J., Wang, Y., Xin, Q., & Zhou, H. (2024). Implementation of an AI-based MRD evaluation and prediction model for multiple myeloma. *Frontiers in Computing and Intelligent Systems, 6*(3), 127–131. https://doi.org/10.54097/zJ4MnbWW

[10] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4*(5), 119–134. https://doi.org/10.69987/JACS.2024.40509

[11] Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

[12] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems, 4*(4), 80–96. https://doi.org/10.69987/JACS.2024.40407

[13] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI Credit Card Clients dataset. *Journal of Advanced Computing Systems, 3*(4), 31–47. https://doi.org/10.69987/JACS.2023.30403

[14] Chen, Y., Zhou, S., & Lin, E. (2025). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. *Journal of Technology Informatics and Engineering, 4*(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542

[15] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-verification reduces hallucination in large language models [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.11495

[16] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*, 625–630. https://doi.org/10.1038/s41586-024-07421-0

[17] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in Neural Information Processing Systems, 30*, 4878–4887. https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[18] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[19] Griot, M., Hemptinne, C., Vanderdonckt, J., & Yuksel, D. (2025). Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications, 16*(1), Article 642. https://doi.org/10.1038/s41467-024-55628-6

[20] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[21] Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine, 30*(9), 2613–2622. https://doi.org/10.1038/s41591-024-03097-1

[22] He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. *Journal of Advanced Computing Systems, 4*(1), 100–120. https://doi.org/10.69987/JACS.2024.40108

[23] He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. *Journal of Technology Informatics and Engineering, 4*(1), 306–324. https://doi.org/10.51903/jtie.v4i1.546

[24] He, S., Nie, J., & Li, C. (2026). Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations. *Journal of Technology Informatics and Engineering, 5*(1), 341–359. https://doi.org/10.51903/jtie.v5i1.548

[25] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48–66. https://doi.org/10.69987/JACS.2023.30404

[26] Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics, 9*, 962–977. https://doi.org/10.1162/tacl_a_00407

[27] Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

[28] Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

[29] Jin, J., Huang, T., & Lu, S. (2024a). Cost-sensitive learning, simulated PU learning, and one-class autoencoding for extreme-imbalance credit card fraud detection. *Journal of Advanced Computing Systems, 4*(6), 64–73. https://doi.org/10.69987/JACS.2024.40605

[30] Jin, J., Huang, T., & Lu, S. (2024b). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI Credit Card Default dataset. *Journal of Advanced Computing Systems, 4*(6), 74–85. https://doi.org/10.69987/JACS.2024.40606

[31] Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering, 4*(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

[32] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[33] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4*(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

[34] Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering, 5*(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

[35] Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design, 3*(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

[36] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerces. *Rule of Law Studies Journal, 2*(2), 105–123. https://doi.org/10.64780/rolsj.v2i2.225

[37] Li, Y. (2024). Findable then explainable: Retrieval-summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65–82. https://doi.org/10.69987/JACS.2024.40706

[38] Li, Y., & Lu, S. (2025). Language-guided feature selection for DDoS and intrusion detection on CICIDS2017. *Journal of Technology Informatics and Engineering, 4*(1), 284–305. https://doi.org/10.51903/jtie.v4i1.531

[39] Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. *International Journal of Graphic Design, 3*(1), 196–215. https://doi.org/10.51903/ijgd.v3i1.3661

[40] Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering, 5*(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

[41] Liu, G., He, S., & Liu, I. (2023). LLM-augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39–57. https://doi.org/10.69987/JACS.2023.30604

[42] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

[43] Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. *Journal of Advanced Computing Systems, 4*(4), 97–111. https://doi.org/10.69987/JACS.2024.40408

[44] Lu, S., & Zhou, D. (2024). TinyLLM-assisted intrusion detection for real-time IoT networks. *Journal of Advanced Computing Systems, 4*(8), 72–87. https://doi.org/10.69987/JACS.2024.40809

[45] Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering, 5*(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

[46] Mi, G., Ye, T., & Wood, D. (2025). A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST. *Journal of Technology Informatics and Engineering, 4*(3), 572–589. https://doi.org/10.51903/jtie.v4i3.492

[47] Nie, J., & Zheng, D. (2024). Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion. *Journal of Advanced Computing Systems, 4*(4), 112–123. https://doi.org/10.69987/JACS.2024.40409

[48] Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. *International Journal of Graphic Design, 4*(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

[49] OpenAI. (2025a). HealthBench [Data set]. Hugging Face. https://huggingface.co/datasets/openai/healthbench

[50] OpenAI. (2025b, May 12). Introducing HealthBench. https://openai.com/index/healthbench/

[51] Romano, Y., Sesia, M., & Candès, E. J. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems, 33*, 3581–3591. https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html

[52] Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self diagnosis and triage: Audit study. *BMJ, 351*, h3480. https://doi.org/10.1136/bmj.h3480

[53] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., . . . Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[54] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., Neal, D., Rashid, Q. M., Schaekermann, M., Wang, A., Dash, D., Chen, J. H., Shah, N. H., Lachgar, S., Mansfield, P. A., . . . Natarajan, V. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine, 31*(3), 943–950. https://doi.org/10.1038/s41591-024-03423-7

[55] Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. *Journal of Technology Informatics and Engineering, 5*(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

[56] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. *Journal of Technology Informatics and Engineering, 4*(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

[57] Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design, 4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

[58] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. *Journal of Advanced Computing Systems, 3*(8), 9–24. https://doi.org/10.69987/JACS.2023.30802

[59] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computational Systems and Applications, 3*(1), 15–30. https://doi.org/10.64229/j6d7fr94

[60] Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

[61] Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine, 7*(1), Article 258. https://doi.org/10.1038/s41746-024-01258-7

[62] Tanno, R., Barrett, D. G. T., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Lau, C., Tu, T., Azizi, S., Singhal, K., Schaekermann, M., May, R., Lee, R., Man, S., Mahdavi, S., Ahmed, Z., Matias, Y., Barral, J., . . . Ktena, I. (2025). Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine, 31*, 599–608. https://doi.org/10.1038/s41591-024-03302-1

[63] Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., Palepu, A., Mustafa, B., Chowdhery, A., Liu, Y., Kornblith, S., Fleet, D., Mansfield, P., Prakash, S., Wong, R., . . . Natarajan, V. (2024). Towards generalist biomedical AI. *NEJM AI, 1*(3), AIoa2300138. https://doi.org/10.1056/AIoa2300138

[64] Wang, B., He, Y., Shui, Z., Xin, Q., & Lei, H. (2024). Predictive optimization of DDoS attack mitigation in distributed systems using machine learning. In Proceedings of the 6th International Conference on Computing and Data Science (pp. 89–94).

[65] Wu, Q., Mi, G., & Wood, D. (2025). Calibration-light subject-independent motor imagery BCI via self-supervised pretraining and Conformer. *Journal of Technology Informatics and Engineering, 4*(1), 239–262. https://doi.org/10.51903/jtie.v5i1.493

[66] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. Emerging Information Science and Technology, 6(2). https://doi.org/10.18196/eist.v6i2.31232

[67] Xin, Q. (2025b). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[68] Xin, Q. (2026a). Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education. *Journal of Applied Artificial Intelligence in Education, 2*(1), 1–19. https://doi.org/10.66053/jaaie.v2i1.348

[69] Xin, Q. (2026b). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit dataset (HELOC-motivated). *Journal of Information and Technology, 14*(2), 215–231. https://doi.org/10.32664/j-intech.v14i02.2228

[70] Xin, Q. (2026c). Host-based intrusion detection with system call sequences: Window localization and forensic narratives. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2), 325–334. https://doi.org/10.28989/avitec.v8i2.3973

[71] Xin, Q. (2026d). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2), 247–264. https://doi.org/10.28989/avitec.v8i2.3974

[72] Xin, Q. (2026e). Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models. Findings. https://doi.org/10.32866/001c.157499

[73] Xin, Q. (2026f). Self-supervised log anomaly detection with LogBERT-style transformers: Full empirical evaluation on a reproducible SynHDFS benchmark. *Journal of Electrical Engineering and Computer Science, 11*(1), 23–35. https://doi.org/10.54732/jeecs.v11i1.3

[74] Xin, Q., Xu, Z., Guo, L., Zhao, F., & Wu, B. (2024). IoT traffic classification and anomaly detection method based on deep autoencoders. In Proceedings of the 6th International Conference on Computing and Data Science.

[75] Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 6233–6251). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.372

[76] Ye, T., Chang, X., & Zhong, E. (2025). Uncertainty-aware breast ultrasound explanation cards: A visual communication framework for image-based AI diagnostic support using BreastMNIST_224. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3701

[77] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109–125. https://doi.org/10.69987/JACS.2024.40308

[78] Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. *International Journal of Graphic Design, 3*(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697

[79] Zhang, J. (2025). From general human activity recognition to volleyball-oriented wearable transfer learning: Cross-dataset evidence from UCI HAR and WISDM for domain adaptation and edge deployment. *Journal of Technology Informatics and Engineering, 4*(1), 263–283. https://doi.org/10.51903/jtie.v4i1.524

[80] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3*(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

[81] Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms. In 2025 5th International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI) (pp. 1012–1018). IEEE. https://doi.org/10.1109/CEI66465.2025.11398480

[82] Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering, 4*(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

[83] Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design, 3*(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

[84] Zhang, Y., & Zhou, Z. (2026). Strategy-aware therapist imitation for emotional support dialogues: A reproducible ESConv study for LLM response control. *Advances in Educational Technology and Psychology, 10*(2), 92–97. https://doi.org/10.23977/aetp.2026.100213

[85] Zhang, Z., Genc, Y., Wang, D., Ahsen, M. E., & Fan, X. (2021). Effect of AI explanations on human perceptions of patient-facing AI-powered healthcare systems. *Journal of Medical Systems, 45*(6), Article 64. https://doi.org/10.1007/s10916-021-01743-6

[86] Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU Trace. *Journal of Technology Informatics and Engineering, 4*(3), 544–571. https://doi.org/10.51903/jtie.v4i3.498

[87] Zhao, S., Ren, Y., & Chang, X. (2026). Profit-aware spot GPU admission control with cost-sensitive loss and evidence-grounded policy memos for AI workload supply-demand matching. *Journal of Technology Informatics and Engineering, 5*(2), 45–59. https://doi.org/10.51903/jtie.v5i2.545

[88] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. *Journal of Advanced Computing Systems, 4*(1), 83–99. https://doi.org/10.69987/JACS.2024.40107

[89] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3*(2), 35–49. https://doi.org/10.69987/JACS.2023.30203

[90] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems, 4*(1), 67–82. https://doi.org/10.69987/JACS.2024.40106

[91] Zhong, Z. S., & Ling, S. (2024a). Improved theoretical guarantee for rank aggregation via spectral method. *Information and Inference: A Journal of the IMA, 13*(3), Article iaae020. https://doi.org/10.1093/imaiai/iaae020

[92] Zhong, Z. S., & Ling, S. (2024b). Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2408.05944

[93] Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

[94] Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering, 5*(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539

[95] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In Y. Li, S. Mandt, S. Agrawal, & E. Khan (Eds.), Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (Vol. 258, pp. 559–567). PMLR. https://proceedings.mlr.press/v258/zhong25a.html

[96] Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. *International Journal of Graphic Design, 3*(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616

[97] Zhong, Z., Zheng, M., Mai, H., Zhao, J., & Liu, X. (2020). Cancer image classification based on DenseNet model. *Journal of Physics: Conference Series, 1651*(1), Article 012143. https://doi.org/10.1088/1742-6596/1651/1/012143

[98] Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696

[99] Zhou, H., & Zhang, K. (2025). News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from 2015–2024 FRED panel. *Journal of Technology Informatics and Engineering, 4*(2), 487–501. https://doi.org/10.51903/jtie.v4i2.540